

Skymapping with OSSE via the Mean Field Annealing Pixon Technique

D. D. Dixon, T. O. Tümer, and A. D. Zych

*University of California
Institute for Geophysics and Planetary Physics
Riverside, CA 92521*

L. X. Cheng

*Astronomy Department
Univ of Maryland
College Park, MD 20742*

W. N. Johnson and J. D. Kurfess

*E. O. Hulburt Center for Space Research
Naval Research Laboratory
Washington, DC 20375*

R. K. Piña and R. C. Puetter

*Center for Astrophysics and Space Science
University of California, San Diego
9500 Gilman Dr.
La Jolla, CA 92093-0111*

W. R. Purcell

*Northwestern University
Department of Physics and Astronomy
2145 Sheridan Rd.
Evanston, IL 60208-3112*

and

Wm. A. Wheaton

*High Energy Astrophysics Group
Jet Propulsion Laboratory, 169-327
California Institute of Technology
Pasadena, CA 91109*

ABSTRACT

We present progress toward using scanned OSSE observations for mapping and sky survey work. To this end, we have developed a technique for detecting pointlike sources of unknown number and location, given that they appear in a background which is relatively featureless or which can be modeled. The technique, based on the newly developed *pixon* concept and mean field annealing, is described, with sample reconstructions of data from the OSSE Virgo Survey. The results demonstrate the capability of reconstructing source information without any *a priori* information about the number and/or location of pointlike sources in the field-of-view.

Subject headings:

1. Imaging with OSSE

The Oriented Scintillation Spectrometer Experiment (OSSE) aboard the Compton Gamma Ray Observatory (GRO) consists of four actively shielded NaI(Tl)-CsI(Na) phoswich detectors (Johnson *et al* 1993). Collimation is provided by tungsten collimators, which define a $3.8^\circ \times 11.4^\circ$ (FWHM) field of view. Each detector is mounted on an independent single-axis pointing system, which allows for instrument pointings which are offset from the spacecraft Z -axis. Complex pointing plans for the OSSE instrument may be effected via the programmable detector orientation system.

The original functional intent of OSSE was to be a “point-and-count” spectrometer, i.e., the detector would be oriented toward a source of interest and counts would be accumulated, with offset pointings providing background estimates for the observation. The ability to program more complex pointing plans, however, raises the possibility of using the non-uniform aperture response to our advantage. By scanning the instrument in steps smaller than the aperture size, and taking scans that overlap in at least two different directions, we can (in principle) use the knowledge of the aperture function and the differential flux measured between detectors to distinguish features at a substantially better resolution than that implied by the aperture size.

A key goal of OSSE scanned observations is to perform sky-survey work, and attempt to detect previously unknown point sources. A separate but related project is to map the low energy galactic γ -ray emission. However, the nature of OSSE scanned observations presents some major difficulties to standard data inversion techniques. Typically, the total number of observations is small ($O(100)$), each with a fairly low signal-to-background ($\sim 0.1\%$). From this, we would like to construct a map of the flux in a

rather larger set of pixels ($O(1000)$), where the pixels are significantly smaller than the aperture size. This is obviously an impossible task for “direct” deconvolution or inversion, and while model fitting is more feasible, it is also undesirable due to the bias introduced by selection of a particular model. We have thus developed a new approach which is highly effective at detecting point sources in an unbiased manner, i.e., with no previous knowledge of their number, location, or strength. This approach begins with the problem phrased as one of deconvolution, and ultimately transforms it to one of model fitting.

2. Direct Deconvolution

OSSE is a *linear* instrument, in that the detected signal is a linear function of the source intensities (this holds true for many types of instruments). Due to the linear nature of OSSE, it is perhaps most “natural” to deal with statistical fluctuations (Poisson noise) in the data by use of *linear least squares* (LLSQ) techniques. For the purposes of this paper, it will be useful to distinguish between two types of LLSQ problems, namely model fitting and deconvolution. By LLSQ *model fitting* we mean the adjustment of the co-efficients in a given linear model to fit a given set of count data (as described, e.g., by Wheaton *et al.* 1995, and references therein). An example would be fitting an OSSE data set to a model consisting of several point sources at known locations, and a linear background model with a finite (small) number of terms of known form. *Deconvolution*, on the other hand, refers to situations in which the model is not known, or perhaps has a known form but a large (in principle infinite) number of terms. An example of deconvolution would be fitting the same OSSE data for observations containing diffuse sources, or with an unknown number of sources at unknown positions.

The essential distinction between the two is

that the former problem involves solving a finite set of linear equations, whereas the latter requires solution of an *integral* equation, e.g.,

$$\bar{D}(y) = \int E(y, x) I(x) dx, \quad (1)$$

where \bar{D} is the expected data, I is the image intensity distribution as a function of spatial coordinates x , and $E(y, x)$ is a known function, essentially the OSSE aperture function in our context, relating the image space X to the data space Y . Solving for $I(x)$ given $D(y)$ and E is a linear inverse problem, of a type common in astronomy.

Discrete LLSQ model fitting methods may be applied to the deconvolution problem if we approximate the linear integral equation (1) by a finite matrix equation:

$$\bar{\mathbf{D}} = \mathbf{E}\mathbf{I}, \quad (2)$$

where now $\bar{D}_i = \bar{D}(y_i)$, $E_{ij} = E(y_i, x_j)$, and $I_j = I(x_j)$. Because of the noise (counting statistics, in the Poisson case), $\bar{\mathbf{D}}$ is never observed directly. The measured quantity is the observed counts \mathbf{D} , which will contain Poisson noise. In astronomical imaging applications, we may try a huge model representing the cosmic source terms, consisting of a source at “every possible location”, that is, an array of pixels spaced closely enough to be a good approximation to $I(x)$ in the integral (1). This approach we have previously called *Direct Linear Algebraic Deconvolution* (DLAD, Dixon *et al.* 1993). Wheaton *et al.* 1995 show that the DLAD estimate coincides with the Poisson Maximum Likelihood (ML) if the pixel array is the same. As DLAD involves solving linear rather than non-linear equations, it is computationally efficient, a point which shall prove crucial in what follows.

Two difficulties arise at once. First, if the data space Y is binned finely to represent the integral faithfully, then for the resulting tiny bins the expected counts $\bar{D}_i = \bar{D}(y_i)$ in data bin

i may be small, so that older Poisson LLSQ (PLLSQ) model-fitting methods fail due to “insufficient statistics” (that is, for small numbers of counts, the Poisson distribution is poorly approximated by a normal distribution). Second, when the pixel array x_j is chosen fine enough, then \mathbf{E} , the discretized matrix representation of $E(y, x)$, becomes highly ill-conditioned or singular. Effective solutions to the first difficulty have been discussed by Wheaton *et al.* 1995. The key point is that the weights for the LLSQ equations should be chosen to be uncorrelated with the data \mathbf{D} itself, effectively ruling out the commonly used approximation $\sigma_i \approx D_i^{\frac{1}{2}}$ (as opposed to $\sigma_i = \bar{D}_i^{\frac{1}{2}}$, which is of course exact). Bearing this requirement in mind, it is possible to construct PLLSQ estimators which are unbiased and otherwise convenient for arbitrarily low \bar{D}_i . With a weighting matrix \mathbf{W} so chosen, the “normal equations” (see references in Wheaton *et al.* 1995) are

$$\mathbf{E}^T \mathbf{W}^2 \mathbf{E} \mathbf{I} = \mathbf{E}^T \mathbf{W}^2 \mathbf{D}, \quad (3)$$

which may be solved for \mathbf{I} , given \mathbf{D} , in the usual way.

The second difficulty has often been summarized by describing the discretization of the linear inverse problem as “ill posed”. We distinguish two aspects of the situation: (a) failure of the numerical mathematical problem, which is essentially an artifact, due to the incapacity of simpler matrix inversion algorithms when faced with ill-conditioned or singular problems (especially *large* problems), and (b) the violent anti-correlation which arises among near-by pixels when the pixel size is \leq angular resolution of the instrument. The algorithmic aspect of the problem is resolved by recourse to singular value decomposition (SVD; see Press *et al.* 1986 for a simple discussion) in the computation. The anti-correlation among pixels is mathematically inescapable, due to the real confusion among nearby points, which the instrument is

unable to resolve clearly. The problem can be partly removed by making use of the *physical* constraint of non-negativity of the source fluxes, as previously discussed in connection with imaging from COMPTEL data (Dixon *et al.* 1993, and Wheaton *et al.* 1993, and more generally by Dixon *et al.* 1996). The non-negativity constraint has been enforced by a variant of the Non-Negative Least Squares (NNLS) algorithm of Lawson and Hanson (1974), and the resulting technique is termed *Constrained Linear Algebraic Deconvolution* (CLAD) (Dixon *et al.* 1996).

Figure 1 shows a map, using data from the OSSE Virgo Survey, made using the CLAD method with a resolution of 2° in the 50–150 keV energy band. The irregularly shaped exterior region bounds the area of significant exposure during the scanned observation. The exposure matrix \mathbf{E} was calculated from the OSSE aperture response, and gives the contribution to each data point from each sky pixel. In this observation, the detectors were stepped at increments of $\sim 1.8^\circ$ for individual scans, with the spacecraft orientation changed $\sim 4^\circ$ between scans. The two strong sources in the FOV, 3C 273 and NGC 4388, appear prominently, but there are several other spurious bright sources, albeit of lower statistical significance. We denote these pixels as “spurious” due to their lack of correspondence with any candidate γ -ray source, as well as their marginal significance (3C 279 is in the field, but no significant feature is associated with it in the map). An important point is that due to the nature of the observations, the reconstructed source intensity doesn’t necessarily have a direct correspondence to its statistical significance. Note that M87 is within the same pixel as NGC 4388, but it is as yet unclear if its contribution to the total flux is significant in this energy range.

While CLAD as illustrated in Figure 1 appears to be a useful tool, the spurious sources are, to say the least, cosmetically unattractive. That

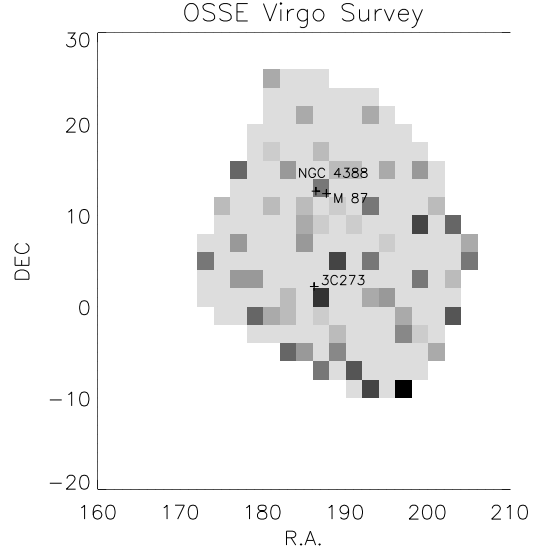


Fig. 1.— Image found generated by solving the DLAD equations, subject to the constraint that the answer be non-negative. Bright sources in the FOV are labelled. Note other bright, spurious sources. The image resolution is 2° .

they are generally non-significant suggests that we could still make a map which fits the data well even if they were somehow eliminated. A more serious problem is that the large number of pixels retained in the model seriously degrades the sensitivity of the maps produced. We are thus led to seek an objective means of eliminating from the model those pixels which are not statistically required by the data. A development based on the *pixon* concept of Puetter and Piña (1993), described in the following section, gives a powerful new image reconstruction algorithm which fulfills this need.

3. Pixon Based Inversion

In a broad sense, the problems we encountered above occurred because we tried to fit too many parameters, i.e., more than were statistically justified by the data. For example, consider a large region in an image that is statistically consistent

with being flat. Ideally, we could represent the information in this region with a single number, such as the average flux. However, since we don't necessarily know *a priori* which regions are flat and which aren't, the standard approach is to cover the FOV with small pixels, and find the flux in each. In the statistically flat region, where there is little coherent information spatial information, but lots of noise, the inversion algorithm quite happily places flux in many of the small pixels to achieve gains (albeit statistically marginal) in the goodness-of-fit (GOF, e.g., χ^2 , likelihood, etc.). The problem, then, is less with the inversion algorithm itself as opposed to the choice of image basis that we supplied to that algorithm.

Many attempts to counteract this problem concentrate on placing some sort of extra constraint (hard or soft) on the form of the solution, and in this way, effectively reducing the number of “free” parameters. A prime example of this is Maximum Entropy (ME) (Skilling 1989). Such approaches are effective to some degree, but often are limited in their realm of applicability, as well as in that they yield biased solutions (Donoho et al 1992). The pixon approach also attempts to reduce the number of free parameters, but does so in a more direct and fundamental manner. For the purposes of this paper, we shall limit somewhat the discussion of the conceptual underpinnings of pixons; for more details, the reader is referred to (Puetter 1994). The basic idea behind a pixon is that it is a sort of flexible pixel (Puetter and Piña 1993), able to change its shape and size. In the pixon approach, a large flat area in the image would not be represented by many pixels containing the same intensity, but rather would be represented by a single pixon, with a single number for the intensity and a few others describing the shape. An image described by pixons would thus require a smaller set of parameters.

The criterion we use in this approach is essentially Occam's razor: the simplest model which

yields an image statistically consistent with the data is the correct model. The next question is how to go about *finding* the simplest model. The starting point is to define a *fundamental model*, consisting of a grid of pixels at the smallest resolution we expect to be able to see, as well as the instrumental response from those pixels. For OSSE, this is represented by the exposure matrix \mathbf{E} , whose columns form the fundamental basis. All pixon bases are simply derived as positive linear combinations of the fundamental basis functions. We represent this linear transformation by the matrix \mathbf{K} , and thus the new model matrix is simply $\mathbf{E}' = \mathbf{E}\mathbf{K}$. Equation 3 then becomes

$$\mathbf{K}^T \alpha \mathbf{K} \mathbf{I}^{(\mathbf{p})} = \mathbf{K}^T \beta, \quad (4)$$

where $\alpha = \mathbf{E}^T \mathbf{W}^2 \mathbf{E}$ and $\beta = \mathbf{E}^T \mathbf{W}^2 \mathbf{D}$, and $\mathbf{I}^{(\mathbf{p})}$ is the *pseudo-image*. Equation 4 may be solved for $\mathbf{I}^{(\mathbf{p})}$ subject to the non-negativity constraint to yield the constrained least-squares coefficients in the pixon basis. To see the actual sky map, we need the representation in the fundamental basis, for which we simply compute $\mathbf{I} = \mathbf{K} \mathbf{I}^{(\mathbf{p})}$.

The reader at this point may wonder why we don't just substitute $\mathbf{I} = \mathbf{K} \mathbf{I}^{(\mathbf{p})}$ into eqn. 4 and solve for \mathbf{I} directly. For that matter, a quick bit of algebra would seem to indicate that all of the \mathbf{K} 's cancel out when solving for \mathbf{I} . However, it is to be noted that direct algebraic solution for $\mathbf{I}^{(\mathbf{p})}$ gives the *unconstrained* least-squares solution. Enforcement of the non-negativity constraint implies that some elements of $\mathbf{I}^{(\mathbf{p})}$ are constrained to be zero. Only those columns of \mathbf{E}' which will yield positive coefficients are used as the basis for the fit, which proves to be the key in making the pixon approach effective. Let us illustrate with a simple problem. We take \mathbf{I} to be two-dimensional and \mathbf{D} to be three-dimensional. \mathbf{E} is thus a 3×2 matrix, which we represent as $[\mathbf{E}_1, \mathbf{E}_2]$, where \mathbf{E}_1 and \mathbf{E}_2 are the 3-vectors forming the fundamental basis. Figure 2 shows the plane spanned by the fundamental basis, where \mathbf{D}_{true} is the true value of \mathbf{D} , which we would measure in the

limit of infinite statistics. Measurement noise, however, will cause \mathbf{D} to be found away from \mathbf{D}_{true} , and generally contain some component perpendicular to the plane. So we perform a least-squares fit to find \mathbf{D}_{LS} , which is simply the orthogonal projection of \mathbf{D} onto the plane. As \mathbf{D}_{LS} lies within the convex region bounded by \mathbf{E}_1 and \mathbf{E}_2 , the components of \mathbf{I} will be positive. Now we form the pixion basis $\mathbf{E}' = [\mathbf{E}_1, \mathbf{E}_1 + \mathbf{E}_2]$, and see that \mathbf{D}_{LS} lies outside the convex region bounded by this basis. The non-negative least-squares estimate $\mathbf{D}_{\text{pixion}}$ is found by projection of \mathbf{D}_{LS} onto the nearest positive subspace (Werner 1990), which in this case is the vector $\mathbf{E}_1 + \mathbf{E}_2$. The pseudo-image is thus $\mathbf{I}^{(p)} = [\mathbf{0}, \mathbf{I}_2^{(p)}]$, giving the image $\mathbf{I} = [\mathbf{I}_2^{(p)}, \mathbf{I}_2^{(p)}]$, which is obviously not the same as the image corresponding to \mathbf{D}_{LS} .

The above exercise, while rather simple, does show precisely why the pixion approach is superior to straightforward constrained ML fitting. The pixion basis, as it is formed of positive linear combinations of fundamental basis functions, will *always* bound a smaller convex region than the fundamental basis. As shown in Figure 2, if this reduced region contains \mathbf{D}_{true} , then the estimate $\mathbf{D}_{\text{pixion}}$ is likely to be closer to \mathbf{D}_{true} than the ML estimate. Occam’s razor has served us well: the fundamental model required two basis functions to perform the fit, but due to the constraint, the pixion model required only one, and yielded a superior estimate. In the context of image reconstruction, Occam’s razor implies that the optimal pixion basis will allow only that level of detail that is statistically justified by the data. Areas in the image which are statistically smooth should be constrained to be smooth in the fit, thus reducing the number of free parameters. As a result, spurious sources are largely done away with, and correspondingly, the flux estimates and sensitivity for actual sources are improved.

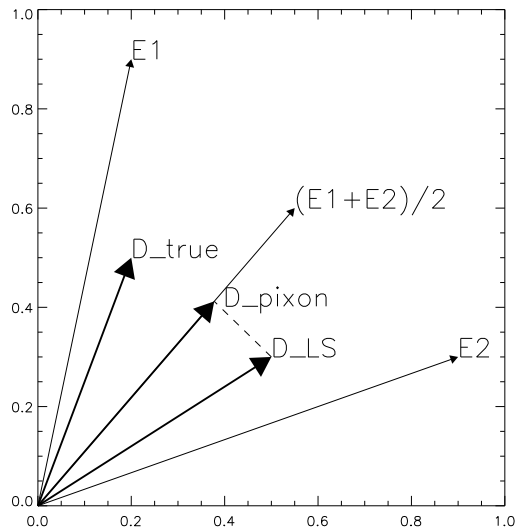


Fig. 2.— A schematic of the effect of the pixion prior. D_{true} is the “noise-free data”, i.e. what we would measure in the limit of infinite statistics. E_1 and E_2 represent fundamental basis functions for the instrument, while $(E_1 + E_2)/2$ is a basis function for a pixion basis. D_{LS} is the constrained least-squares estimate for the fundamental basis, while D_{pixion} is the constrained least-squares estimate for the pixion basis. Note that D_{pixion} requires only one basis function as opposed to two for D_{LS} , and further that D_{pixion} is closer to D_{true} .

4. Finding the optimal pixion basis

While the advantages of using the optimal pixion basis for image reconstruction are obvious, what is not obvious is how one goes about finding it. The set of possible basis functions consists of *all* positive linear combinations of the fundamental basis functions. Further, there exists no *a priori* systematic way of ordering the possible ways of choosing a basis, i.e., the only way compare the merits of two different bases is to actually perform the fit for each one. We must find some way of systematizing the search procedure.

The first step is to choose some set of potential pixion shapes. This eliminates the possibility of being absolutely optimal, but greatly shrinks the set of possible basis functions. Following Puetter and Piña (Puetter and Piña 1993), we have implemented a *fuzzy pixion basis*. Here, the pixion boundaries are not sharply defined, and the pixions are allowed to overlap. For simplicity, we shall use pixions which are circularly symmetric, and thus describable by a single parameter. For the pixion shape we have chosen, there will be associated with each pixel in the image a length δ_j , called the pixion scale, which defines the size of the pixion. The pixion shape function we have chosen is an inverted paraboloid (Puetter 1994), and thus the pixion transform matrix \mathbf{K} is defined as

$$\begin{aligned} F_{jk} &= \begin{cases} 1 - \frac{d_{jk}^2}{\delta_j^2}, & d_{jk} \leq \delta_j, \\ 0, & d_{jk} > \delta_j, \end{cases} \\ V &= \sum_j F_{jk}, \\ K_{jk} &= \frac{F_{jk}}{V}, \end{aligned} \quad (5)$$

where d_{jk} is the distance between image pixel j and pseudo-image pixel k , and we have a one-to-one correspondence between the image and pseudo-image pixels. Further, we restrict δ_j to some finite set of values. The implementation used for this paper utilizes two values, a small one $\delta^{(1)}$ (usually one pixel) for pointlike sources, and a large one $\delta^{(2)}$ (roughly the size of the image space) for modeling diffuse background emission.

Given the above restrictions on K , we now have a well-defined set of potential basis functions. However, this set is still forbiddingly large. With two pixion scales and an image space of dimension J , the number of possible bases is 2^J , thus a brute force search is infeasible for all but very small problems. The problem of finding an optimal pixion basis is thus one of combinatorial (rather than functional) minimization. Practical techniques for attacking such problems are gener-

ally heuristic in nature, and promise only a near-optimal solution. For OSSE sky survey work, we anticipate looking for point sources within an otherwise roughly uniform background, implying that we really only need two pixion sizes, one small (for point sources) and one large (for the background). The choice of two disparate types of basis functions leads naturally to an optimization heuristic called *mean field annealing* (MFA) (Reeves 1993). Annealing algorithms derive from the statistical mechanical analog of annealing a material from an initially disordered state, where one begins at a high temperature and slowly cools the material. Consider the case of a ferromagnetic material. In the presence of a uniform external field, each atom in the material can have one of two possible spins (up or down), and the configuration space of the magnet is simply all possible combinations of ups and downs. The configuration corresponding to a defect-free magnet occurs at the global minimum of the Hamiltonian, where all of the spins point in the same direction. However, the Hamiltonian also has many local minima, corresponding to the formation of domains, where groups of spins in different regions of the magnet point in different directions. To anneal a magnet, one begins at a temperature high enough that kT greatly outweighs the interaction energy. At high temperature, the spins will be oriented randomly, and on average half will be up and half will be down. If the system is cooled too quickly, it “quenches”, becoming trapped in a local minimum. Slow cooling, however, allows the system to “jump out” of local minima via thermal fluctuations. By lowering the temperature in steps, and allowing equilibrium to be reached at each step, we can arrive at a nearly defect-free magnet in the low temperature limit.

The problem of finding the pixion basis is similar. Here, the configuration space is all possible assignments of pixion sizes to the pixels of the fundamental basis. With two possible pixion

sizes, the problem is closely analogous to that of the magnet, where we tried to assign one of two possible values of spin. The Hamiltonian we choose consists of two parts: the χ^2 , and a function $\sum_j s_j$, where s_j is a linear function of δ_j , and is taken to be 1 at the smallest value of δ_j , and 0 for the largest value. This second function acts as a penalty against complexity, utilizing the fact that smaller pixons (more detail) generally imply more parameters in the model. The particular functional relationship of s_j to δ_j isn't so important, for as we'll see below, it is only *changes* in s_j that will be relevant. We then write the following Hamiltonian for the problem:

$$H(\delta) = \gamma\chi^2 + \sum_j s_j, \quad (6)$$

with χ^2 evaluated for the result of a fit using a particular set of pixon scales δ . This Hamiltonian represents a balance between finding the simplest model (all big pixons) and satisfying the χ^2 , with the constant γ weighting the relative importance of the two terms. We then define a parameter T analogous to temperature, and write the Boltzmann distribution

$$p(\delta, T) = \frac{e^{-H(\delta)/T}}{Z}, \quad (7)$$

where Z is the partition function, $Z = \sum_{\delta} e^{-H(\delta)/T}$, with the sum performed over all possible configurations δ .

Our ultimate goal is to find the average values $\langle \delta \rangle$ at a given temperature for the specified Hamiltonian and Boltzmann distribution. Clearly, this calculation cannot be carried out exactly for an interesting number of pixels, as it requires calculating eqn. 7 over 2^J configurations. We note that for T constant, the equilibrium energy is equal to the average energy. Define H_{ij} as the average energies for states with δ_j held fixed at one of the two allowed values $\delta_j^{(i)}$, i.e., $H_{ij} = \langle H(\delta) \rangle|_{\delta_j = \delta_j^{(i)}}$. We then approximate the average value of δ_j at equilibrium for a given

temperature eqn. 7 as

$$\begin{aligned} \langle \delta_j \rangle &= \frac{\delta_j^{(1)} \exp(-H_{1j}/T) + \delta_j^{(2)} \exp(-H_{2j}/T)}{\exp(-H_{1j}/T) + \exp(-H_{2j}/T)} \\ &= \frac{\delta_j^{(1)} + \delta_j^{(2)} \exp[(H_{1j} - H_{2j})/T]}{1 + \exp[(H_{1j} - H_{2j})/T]}. \end{aligned} \quad (8)$$

where the mean or effective field at j is $\Phi_j = H_{1j} - H_{2j}$. This seems easy enough. However, calculation of H_{ij} still requires summation over *all* possible configurations. We thus make the additional approximation (Van den Bout and Miller 1990), that $\langle H(\delta) \rangle \simeq H(\langle \delta \rangle)$, i.e., the average value of the energy is equal to the Hamiltonian evaluated at the average values of δ (for the reader who finds this procedure a bit ad hoc and suspicious, we shall investigate it more deeply in the following section). At each step in the algorithm, then, $\langle \delta_j \rangle$ is calculated from $H(\langle \delta \rangle)|_{\delta_j = \delta_j^{(i)}}$, with the values in $\langle \delta \rangle$ being taken from the previous step, leading to the MFA algorithm (Van den Bout and Miller 1990):

1. $T = T_0 \equiv$ initial temperature
2. do until saturation
 - (a) do until equilibrium
 - i. choose a random pixel j
 - ii. set $\delta_j = \delta_j^{(1)}$
 - iii. find the CLAD solution for this basis
 - iv. Compute H_{1j} for that solution
 - v. set $\delta_j = \delta_j^{(2)}$
 - vi. find the CLAD solution for this basis
 - vii. Compute H_{2j} for that solution
 - viii. compute $\langle \delta_j \rangle$ from eqn. 8
 - (b) $T = cT$

“Equilibrium” is defined as the point where the δ_j 's are no longer changing significantly at a particular temperature (in practice, this seems to require only one or two iterations). “Saturation”

is the point where the δ_j 's all take on one of the possible predefined values, in this case, $\delta^{(1)}$ or $\delta^{(2)}$. The constant c represents the cooling rate, and must be chosen carefully. If it is too small, the system quenches, and the low temperature limit produces a solution that is in a non-optimal local minimum. If c is too large, the algorithm takes too long to complete. Nominally, the initial temperature may be chosen such that the system is initially maximally disordered, i.e., $\delta_j = (\delta^{(1)} + \delta^{(2)})/2$. In practice, one can choose a somewhat smaller temperature, where some degree of structure (though not too much) is present in the δ_j 's. The constant γ is also of importance, as overweighting of the χ^2 gives spurious sources, and underweighting suppresses weak sources. Note also from eqn. 8 that we only use the *difference* between H_{1j} and H_{2j} . Since we modify only the δ_j for a particular j while holding the others constant, $H_{1j} - H_{2j} = \Delta\chi^2 - \Delta s_j$. Given the definition of s_j , Δs_j is simply 1.

We note that this algorithm requires a good deal of solving of the least squares problem for different bases, which may seem to impose a large computational burden. However, the variant of NNLS we have developed mitigates this to a great extent. Partially, this is because we are solving the least squares problem within a linear algebraic context, which is much more computationally efficient than non-linear minimization. Another contributing factor is our variant of NNLS, which takes advantage of the fact that only some subset of the parameters is unconstrained, and uses a QR factorization updating scheme rather than re-solving the entire problem when only a single δ_j is changed. This algorithm will be described in a subsequent paper; a brief description is given here. The QR -decomposition algorithm factors a matrix into an orthonormal part (Q) and an upper triangular part (R) (Gill, Murray, and Wright 1991). A useful property of QR factorization is that rank one updates (e.g., addition of a column to the original matrix) are relatively

easy to perform (Gill, Murray, and Wright 1991). The original Lawson and Hanson implementation of NNLS takes advantage of this (Lawson and Hanson 1974). Within the pixon framework, addition/removal of a particular pixon to/from the active set is also a rank one update. Further, this action generally doesn't modify the constraint set a great deal. By using the results from a previous NNLS step, and applying the QR -updating, we generally save a good deal of time over redoing the entire NNLS procedure. The result is that for J pixels, N possible pixon sizes, L temperature steps, and an average of M unconstrained parameters, the algorithm performs like $O(JLNM^2)$.

5. Further discussion of the algorithm

In this section, we present a more detailed look at the inner workings of the mean field annealing algorithm described above. The “proof is in the pudding” reader may wish to skip this section on first reading, and move directly to the section on the application of the algorithm to OSSE data.

The above derivation of the mean field equations (following that in (Van den Bout and Miller 1990)), while based on seemingly reasonable assumptions, does not make much rigorous contact with statistical mechanics. To make this connection clear, we will present the derivation given in (Reeves 1993), which is a standard mean field approximation procedure. We begin with a set of J “spins”, s , which are allowed to take on the values 0 and 1. The interactions between these spins is described by a Hamiltonian or energy function, $H(s)$. Correspondingly, for a system which is Boltzmann distributed at temperature T , the partition function Z is given by

$$Z = \sum_{[s]} e^{-H(s)/T}, \quad (9)$$

where the sum is taken over all possible configurations of s . Next, we replace s with continuous

variables v :

$$Z = \sum_{[s]} \int d[v] e^{-H(v)/T} \prod_j \delta(s_j - v_j), \quad (10)$$

which, via the presence of the δ -function, is the equal to the previous expression. The function $H(v)$ now represents an effective energy in terms of the new variables. We then Fourier expand the δ -functions in conjugate variables u , obtaining

$$Z = \sum_{[s]} \int d[v] \int d[u] e^{-H(v)/T} \prod_j e^{u_j(s_j - v_j)}, \quad (11)$$

and carry out the sum over s :

$$Z \propto \int d[v] \int d[u] e^{-H(v)/T - \sum_j u_j v_j + \sum_j \log(1 + e^{u_j})}. \quad (12)$$

We have thus rewritten the partition function entirely in terms of the new variables, and as yet have made no approximation. Next, the saddlepoint approximation is made, which approximates the value of the integrand by its maximum value. The point where the integrand takes on its maximum value is given by

$$\begin{aligned} u_j &= -\frac{1}{T} \frac{\partial H(v)}{\partial v_j}, \\ v_j &= \frac{1}{1 + e^{u_j}}. \end{aligned} \quad (13)$$

Combining these gives the implicit equation for v_j :

$$v_j = \left(1 + e^{-\frac{1}{T} \frac{\partial H(v)}{\partial v_j}} \right)^{-1}. \quad (14)$$

The question immediately arises as to the interpretation of the new variables v , for at first glance, eqn. 14 bears little resemblance to something like eqn. 8. For many optimization problems, such as graph partitioning (Van den Bout and Miller 1990), the form of H is such that $\partial H / \partial v_j = \Phi_j$, the mean field at spin j . In this case, v_j in eqn. 14 takes the form of a thermal

average of s_j , evaluated in the mean field approximation. The Hamiltonian in our application, however, is not so simple. In this case, the exponent in eqn. 14 is given by

$$\gamma \frac{\partial \chi^2}{\partial v_j} + 1, \quad (15)$$

where the functional dependence of χ^2 on v is given via the relation $\delta_j = \delta^{(1)}(1 - v_j) + \delta^{(2)}v_j$. Given that we evaluate χ^2 *after* solving the constrained least squares problem in the basis associated with δ , it is not at all clear that one can obtain a functional form for its partial derivatives. Even if we were to make the reasonable assumption that small perturbations in δ do not change the constraint set, the expression for the partial derivative is quite complicated, and difficult to evaluate numerically. An obvious remedy is to make a numerical approximation of the derivative,

$$\frac{\partial \chi^2}{\partial v_j} \simeq \frac{\Delta \chi^2}{\Delta v_j}, \quad (16)$$

with $\Delta v_j = v_j^{(2)} - v_j^{(1)}$, and $\Delta \chi^2$ the difference in χ^2 values evaluated at $v_j^{(2)}$ and $v_j^{(1)}$. Substituting the approximate derivative into eqn. 14, setting $\Delta v_j = 1$, and doing a little algebra, we obtain

$$v_j = \frac{e^{-H_{2j}}}{e^{-H_{1j}} + e^{-H_{2j}}}, \quad (17)$$

In this expression, the interpretation of v_j as the thermal average of s_j within the above approximations, is obvious, and as such, leads directly to eqn. 8.

Another question which arises is whether or not the mean field approximation is even valid for the problem at hand. Methods exist in statistical mechanics for deciding this (Pfeuty and Toulouse 1977), but are difficult to apply in a general manner, given the nature of H in our problem. Generally speaking, one expects the approximation to be reasonable in the case where the number of “spins” is large, so that individual interactions

are small compared to the mean field, and/or when the dimensionality of the system is large, such that each “spin” has many neighbors. For the image reconstruction problem, we would expect the first condition to be satisfied when the FOV is a fair amount larger than the extent of the instrument response, such that there exist many groups of pixels which are effectively decoupled. The second condition is more subtle. The dimensionality of our system is nominally 2, but, when we are attempting to superresolve the image, i.e., the pixel size is smaller than the extent of the instrument response, the “forces” in the problem are effectively long range, as spatially separated pixels will be correlated via the response. For long range forces, the effective dimension of the system is larger than the geometric dimension (Pfeuty and Toulouse 1977). The precise calculation of this number for the image reconstruction problem is difficult, and problem dependent. Regardless, it should be clear that for the case of small pixels, each “spin” will feel the influence of many nearby “spins”.

Now, we have obviously made just about the coarsest approximation possible for $\partial H/\partial v_j$, but there are some decided benefits to the particular choice we have made. First, as we saw above, it allows us to recover the interpretation of v as an approximate thermal average. Closely related is the property that for the approximation we have chosen, as $T \rightarrow 0$, v_j (and thus $\langle \delta_j \rangle$) takes on one of its two extremal values, which was really our goal in the first place. Due to the complicated nature of $H(v)$, this won’t necessarily occur for the exact solution of eqn. 14 in the low T limit. This property is of some importance in sky survey work, where we don’t really want sources showing up as different sized pixons depending on the statistics of the dataset.

The final benefit has to do with the method of solving for v_j . As eqn. 14 is implicit in v_j , direct solution is impossible. A simple way of solving

for v_j is that of the algorithm described above, which in terms of eqn. 14 is just the iteration

$$v_j^{(n+1)} = \left(1 + e^{-\frac{1}{T} \frac{\partial H(v^{(n)})}{\partial v_j^{(n)}}} \right)^{-1} = f(v^{(n)}). \quad (18)$$

The solution of eqn. 14 is represented by a fixed point (of which there is likely more than one) of the mapping defined in eqn. 18. Now, as $f(v^{(n)})$ of eqn. 18 necessarily lies between 0 and 1, we are guaranteed the existence of a fixed point. We are *not*, however, guaranteed of convergence to that fixed point. The stability of the fixed point is basically determined by the eigenvalues of the Jacobian of the map in the neighborhood of the fixed point (actually, the analysis is somewhat more complicated in the serial updating scheme which we are using, but the flavor is the same (Reeves 1993)). If the eigenvalues λ_i are such that $-1 < \lambda_i < 1$, the fixed point will be stable, and we will have convergence. If not, other undesirable behavior will occur. Though a full eigenvalue analysis is clearly out of the question, numerical experiments have shown that using “better” approximations for $\partial H/\partial v_j$ will often result in behavior which is oscillatory, or even chaotic. In our coarse approximation, however, $\Delta v_j = 1$, and $0 \leq f(v) \leq 1$, so $\partial v_j^{(n+1)}/\partial v_j^{(n)} \leq 1$. Though we don’t prove that this implies stability of the fixed point, we do note that the nature of the problem is such that the Jacobian will likely be diagonally dominated (i.e., pixons which are well separated have little effect on each other w.r.t. v or δ), and so we would expect our coarse approximation to have some stabilizing effect; and in fact, in all of our applications of the algorithm, convergence has always been achieved using this approximation.

Though we have argued for the proposed algorithm’s effectiveness in the case of *two* pixon sizes, it should be fairly clear that it won’t be very effective for three or more. In the above, we see that $H(v)$ is essentially replaced by a linear

approximation, which is fine for capturing the average behavior between two extremal pixion sizes. However, with an intermediate pixion size, one expects that, in some cases, H will have a significant minimum in between the two extremal values, and the linear approximation fails to capture this. Thus the extension of the MFA algorithm to > 2 spin states as suggested in (Van den Bout and Miller 1990) is not appropriate for the problem at hand. We suggest that > 2 pixion sizes might be accommodated within the MFA framework via a higher order approximation for H , but leave this as an avenue for future research.

6. Application to OSSE data

To test the effectiveness of our MFA pixion algorithm, we applied it to the same dataset used to compute Figure 1. For this purpose, we set the initial temperature $T_0 = 3.$, $\gamma = 0.01$, and $c = 0.3$. The computation required approximately 15 minutes of CPU time on a VAXstation model 4000-90 to compute the image for 209 pixels. The result is shown in Figure 3, and is obviously far superior to the constrained least squares result. Most notably, *none* of the spurious sources seen in Figure 1 are found by the MFA algorithm, and the two expected sources stand out clearly. An apparent broad gradient is also noticeable; this is a known background effect for OSSE scanned observations, which we have intentionally not removed to show the utility of the pixion approach. For comparison with a more established technique, Figure 4 shows a Maximum Entropy reconstruction of the same dataset. The suppression of spurious sources by the MFA pixion algorithm is not only aesthetically pleasing, but also implies that more confidence can be placed in the flux estimates of the real sources. Another measure of the success of the MFA algorithm is the number of basis functions required by the fit. The CLAD fit, using the NNLS algorithm, required 56 basis vectors. The pixion image required only 4. It should be

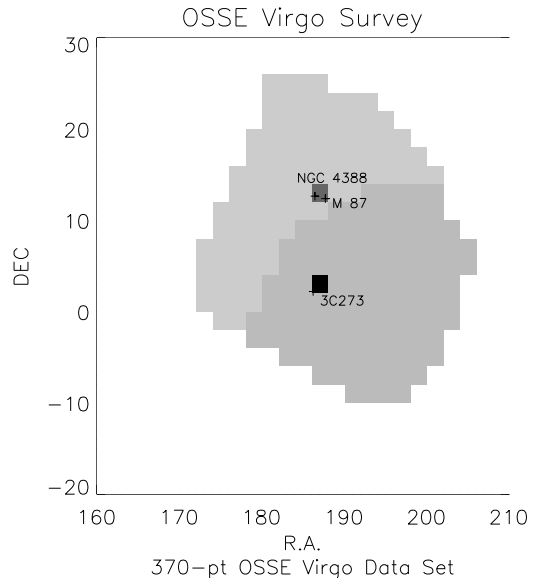


Fig. 3.— Image generated using the MFA pixion algorithm to find a near-optimal pixion basis. Note that the strong sources stand out, but that the spurious peaks common in Figure 1 have been suppressed. The image resolution is 2° .

emphasized that the image in Figure 3 was generated without any *a priori* bias given to those pixels containing real sources. All pixels are treated equally by our algorithm, with no assumptions being made about source locations. Further results for the Virgo survey are presented in a separate paper (Kurfess *et al* 1997).

7. Future directions

The MFA pixion algorithm proves to be remarkably successful, despite certain limitations which will be addressed by future research. Most notably, there currently exists no rigorous way to estimate the parameters T_0 , c , and γ . For application to OSSE data, we have made a heuristic estimate of $\gamma = 1/\Delta\chi^2(3\sigma)$, where $\Delta\chi^2(3\sigma)$ is the change in the χ^2 required to produce a 3σ change w.r.t. the χ^2 distribution where the number of degrees of freedom is calculated from

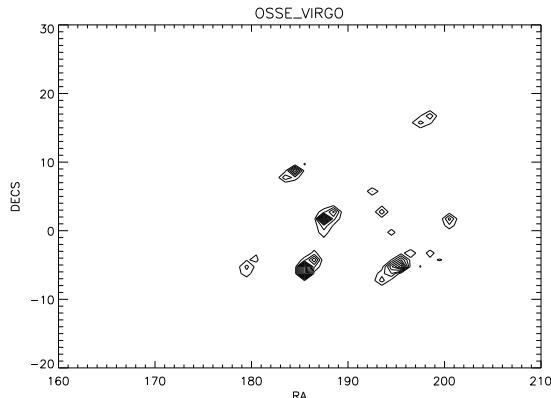


Fig. 4.— The same data as in Figure 3 reconstructed with Maximum Entropy. Apparent sources other than those which are known to occupy the region are of marginal significance, and we attribute them to the scan dependent background effect, which appears as a broad gradient in Figure 3.

the number of data points. The idea is that any addition of further detail to the model which produces less than a 3σ improvement to the χ^2 will be outweighed by the pixon penalty of $\Delta s_j = 1$. From there, we may further refine our estimate of γ by examination of the final χ^2 value. If the χ^2 falls outside the desired confidence interval, γ should be lowered to admit more parameters in the final model. The other two parameters, T_0 and c were essentially guessed. One would like to minimize T_0 , so as not to spend time computing iterations which produce no structure in the image. By examining results from the first iteration, one can tell if T_0 was chosen too large or too small. If too large, the δ_j 's will show no structure, all having essentially the same value. If too small, some of the δ_j 's will saturate to one of the predefined values. Somewhere in between these two extremes lies the correct T_0 , where some (but not too much) structure is apparent, but at this point, judgements of how much structure is appropriate are basically subjective. Similarly, a small value of c will hasten the convergence of the algorithm, but a value too small will lead to

a solution in a non-optimal local minimum. For both T_0 and c , it is best to err on the side of caution and *overestimate* their values. This leads to a somewhat larger computational burden, but also increases confidence that the near-global optimum will be found.

Despite this seemingly large amount of arbitrariness, it turns out that the results from the MFA pixon algorithm are essentially constant over a large range of parameters. Experience has shown that even quite bad estimates of T_0 , c , or γ still lead to excellent results. We have changed the weighting factor γ over as much as two orders of magnitude, and at worst have seen only a few low significance spurious sources, which still represents a vast improvement over previous methods. Local minima obtained when T_0 or c are underestimated generally show sources shifted by one or two pixels, but statistically important detail is never omitted completely (provided γ isn't horribly underestimated).

8. Conclusions

The ability to create skymaps using data from OSSE scanned observations has been demonstrated. For this purpose, we have generated a new image reconstruction algorithm which utilizes the concept of the *pixon*. The driving force behind the pixon concept is simply Occam's razor, where we attempt to find the simplest possible basis for the reconstruction that yields an acceptable value for the χ^2 . As the problem of finding such a basis is combinatoric in nature, we have employed an approximate technique, mean field annealing, which gives a near optimal solution with reasonable computational requirements. The MFA pixon algorithm has been successfully applied to scanned data from the OSSE Virgo Survey, producing skymaps at resolution substantially smaller than the size of the OSSE aperture, while suppressing spurious sources common in pure maximum likelihood re-

constructions. Further, our algorithm has been shown to find the nearly minimum basis for the reconstruction, implying smaller errors on source flux estimates. The algorithm is most effective for the case where the source distribution consists of some number of highly localized sources in an otherwise slowly varying “background”. We see that the MFA pixon algorithm occupies a place somewhere between standard deconvolution algorithms and model fitting algorithms, as it essentially transforms the former problem into the latter.

One final note: the algorithm we have described has a somewhat different flavor than commonly used regularized inversion schemes. Techniques such as Maximum Entropy estimate the *image* by finding the minimum of the sum of two (or more) functions of the image parameters. One of these functions specifies the goodness-of-fit, while the other tends to have the property of suppressing oscillatory behavior in the solution. The MFA pixon algorithm, however, clearly is *not* estimating the image, but rather δ . In this case, the χ^2 plays the role of a constraint, with γ acting as a Lagrange multiplier. From these δ , the image is obtained via a simple constrained least squares estimate. So, despite the fact that it appears that we’ve reduced the number of parameters in the problem, we’ve actually just changed the parameter set. The effectiveness of the technique is at least due in part to the fact that between the χ^2 and the non-negativity requirement on the image, that the total number of constraints is large, and the problem thus highly overdetermined.

DDD wishes to thank J. Wudka for helpful discussions concerning mean field theory. Part of the work described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

This work was funded by NASA CGRO Grants 5-2044 and 5-2825.

REFERENCES

- Dixon, D. D., Buchholz, J., O’Neill, T. J., Tümer, O. T., White, R. S., Zych, A. D., Wheaton, W. A., 1993, *Proceedings of the 23rd International Cosmic Ray Conference*, Calgary, volume 2, 564
- Dixon, D. D., Bhattacharya, D., O’Neill, T. J., Tümer, O. T., White, R. S., Zych, A. D., and Wheaton, Wm. A., 1996, *ApJ*, **457**, 789
- Donoho, D. L. et al, 1992, *J. Roy. Statist. Soc. B*, **54**, 41–81.
- Gill, P. E., Murray, W., Wright, M. G., 1991, *Numerical Linear Algebra and Optimization*, (Redwood City: Addison-Wesley)
- Johnson, W. N., Kinzer, R. L., Kurfess, J. D., Strickman, M. S., Purcell, W. R., Grabelsky, D. A., Ulmer M. P., Hillis, D. A., Jung, G. V., and Cameron, R. A., 1993, *ApJS*, **86**, 693
- Kurfess, J. D. *et al*, 1997, in preparation
- Lawson, C. L. and Hanson, R. J., 1974, *Solving Least Squares Problems*, (New York: Prentice-Hall)
- Pfeuty, P. and Toulouse, G., 1977, *Introduction to the Renormalization Group and to Critical Phenomena*, (Chichester: John Wiley & Sons)
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T., 1986, *Numerical Recipes*, (New York: Cambridge University Press)
- Puetter, R. C. and Piña, R. K., 1993, *SPIE Vol. 1946 Infrared Detectors and Instrumentation*, 405
- Puetter, R. C., 1994, *SPIE Vol. 2302 Image Reconstruction and Restoration*, 112

- Reeves, C. R., 1993, in *Modern Heuristic Techniques for Combinatorial Problems*, C. R. Reeves, Ed., (New York: John Wiley and Sons), 197
- Skilling, J., 1989, in *Maximum Entropy and Bayesian Methods*, J. Skilling, Ed., (Dordrecht, Netherlands: Kluwer Academic Publishers), 45
- Van den Bout, D. E. and Miller, T. K. III, 1990, *IEEE Trans. Neural Networks*, 1, 192
- Werner, H. J., 1990, *Linear Algebra and its Applications*, 127, 379
- Wheaton, Wm. A., Dunklee, A. L., Jacobson, A. S., Ling, J. C., Mahoney, W. A., Radocinski, R. G., 1995, *ApJ*, 438, 322
- Wheaton, Wm. A., Dixon, D. D., O'Neill, T. J., Tümer, O. T., White, R. S., and Zych, A. D., 1993, *Proceedings of the 23rd International Cosmic Ray Conference*, Calgary, volume 2, 560